

Diagnostic Plots For Robust Multivariate Methods

Greet Pison and Stefan Van Aelst

April 24, 2003

Abstract

Recently robust techniques for multivariate statistical methods such as principal component analysis, canonical correlation analysis and factor analysis have been constructed. In contrast to the classical approach, these robust techniques are able to resist the effect of outliers. However, there does not yet exist a graphical tool to identify in a comprehensive way the data points that do not obey the model assumptions. Our goal is to construct such graphics based on empirical influence functions. These graphics not only detect the influential points but also classify the observations according to their robust distances. In this way the observations are divided in four different classes which are regular points, non-outlying influential points, influential outliers, and non-influential outliers. We thus gain additional insight in the data by detecting different types of deviating observations. Some real data examples will be given to show how these plots can be used in practice.

Key words: Empirical influence function; Graphics; Outliers; Robust multivariate methods; Robust Distances.

1 INTRODUCTION

In this paper we construct a graphical tool to identify observations that deviate from an assumed multivariate model. We consider principal component analysis, canonical correlation analysis and factor analysis. Throughout the paper we assume that these methods are based on the correlation matrix (see e.g. Johnson and Wichern 1998). The proposed diagnostic plots are based on robust estimates of the model parameters and empirical influence functions. Recently robust methods for multivariate analysis have been introduced for e.g. principal component analysis (Croux and Haesbroeck 2000), canonical correlation analysis (Croux and Dehon 2001) and factor analysis (Pison et al. 2003). The robustness of the methods has been investigated by means of the influence function. These papers also present empirical influence functions that can be used in the finite-sample case to determine the influence of each observation on the parameter estimates. However, since these multivariate models contain several (high dimensional) parameters, we obtain several empirical influence functions of the same dimensions. Hence, as noted by Shi (1997), a graphical tool

Greet Pison is Statistician, Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium. Stefan Van Aelst is Statistician, Department of applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium.

to represent this information in a comprehensive way is desirable. The goal of these plots is thus to show what effect each of the observations has on the estimated model parameters.

In multivariate analysis (based on the correlation matrix) it can easily be seen that not all observations with large robust distance (outliers) are necessarily influential points that are harmful in the analysis. This is very similar to regression analysis where a large outlier in the carriers can have a very small standardized residual and therefore is called a "good leverage point." In regression analysis a diagnostic plot was proposed by Rousseeuw and van Zomeren (1990) to easily identify vertical outliers, good leverage points and bad leverage points. Similarly, we now construct a graphical tool for multivariate statistical methods. These plots can be very useful in exploratory data analysis or model building stage to assess the quality of a fit and detect unusual data points. The detection of influential points and outliers will be based on two marginal tests that assume under the null hypothesis respectively that the data contain no influential points and no outliers.

In Section 2 we focus on empirical influence functions and explain which version best serves our goal. Here we also introduce the robust estimators of multivariate location and scatter used throughout the paper. In Section 3 we construct the diagnostic tool for principal component analysis. Section 4 focuses on canonical correlation analysis while factor analysis is considered in Section 5. Simple generated data will be used to explain the interpretation of the plots. Real data examples will be analyzed to illustrate the practical use of the plots. We conclude with a discussion in Section 6.

2 EMPIRICAL INFLUENCE FUNCTION

First we introduce the concept of statistical functionals. Suppose we have a p -dimensional estimator T_n which is defined for any sample size n . A statistical functional corresponding to the estimator T_n is a map T which maps any p -variate distribution G on (a subset of) \mathbb{R}^p such that $T(\hat{F}_n) = T_n$ for any possible empirical distribution function \hat{F}_n . The influence function of T at the distribution F can be defined as a Gâteaux derivative of T evaluated at F . The influence function measures the effect on the functional of an infinitesimal contamination at a certain point. See Hampel et al. (1986), Fernholz (1983) for more information on functionals and influence functions. Throughout the paper we will (abusively) associate the influence function with an estimator when we mean the influence function of the functional corresponding to the estimator.

The development of methods to measure the influence of observations on a statistical analysis started for one-dimensional statistics (see Gnanadesikan 1977). The biggest effort has been made in linear regression. Overviews of these developments can be found in e.g. Cook and Weisberg (1982), Fox (1991), Atkinson and Riani (2000). In multivariate analysis influence measures have been discussed by Campbell (1978) in discriminant analysis, Critchley (1985), Tanaka (1988) and Shi (1997) in principal component analysis, Romanazzi (1992) in canonical correlation analysis and by Tanaka and Odaka (1989) in factor analysis. Most of these methods use the (deleted) empirical influence function. The empirical influence function (EIF) is obtained from the influence function by replacing the unknown true distribution F by the empirical distribution \hat{F}_n , so population parameters (e.g. $\mu(F)$ and $\Sigma(F)$) are estimated by their empirical counterparts ($\mu(\hat{F}_n)$ and $\Sigma(\hat{F}_n)$).

To detect the most influential observations, we substitute robust parameter estimates in the influence functions of the classical functionals for the multivariate model as proposed by Lu et al. (1997) and Pison et al. (2002). This is indeed the only EIF that detects all the influential points. Substituting the classical estimates in the influence functions can mask influential outliers because these outliers already influenced the estimates. The standard empirical influence function can thus fail to detect influential outliers due to the masking effect and the deleted empirical influence function cannot resolve this masking if there are clusters of outliers in the data. On the other hand, the influence functions of the robust estimators (see e.g. Jaupi and Saporta 1993) cannot be used to detect the influential outliers because the robust method downweights all outliers, resulting in a small influence. However, robust estimates do not suffer from the masking effect, so they are close to the true parameter values that were to be estimated. Therefore, the influence of observations on the multivariate model can be measured by the empirical influence functions of the classical estimators based on these robust parameter estimates.

Several robust estimators of location and scatter have been proposed in the literature such as M-estimators (Maronna 1976), the Minimum Volume ellipsoid and Minimum Covariance Determinant (MCD) estimators (Rousseeuw 1984, 1985), S-estimators (Davies 1987, Rousseeuw and Leroy 1987) and CM-estimators (Kent and Tyler 1996). See Maronna and Yohai (1998) for an overview. We propose to use the one-step reweighted MCD estimator or S-estimators. These are highly robust estimators with bounded influence function and positive breakdown value that are easy to compute.

Suppose we have a p -dimensional dataset of size n . Among all subsets containing a fraction γ ($0 < \gamma < 1$) of the data, the MCD selects the subset with the smallest determinant of its covariance matrix. The MCD location (t_n^0) is then the empirical mean of that subset, and the MCD scatter estimator (C_n^0) is a multiple of its covariance matrix. The MCD can be computed efficiently with the algorithm of Rousseeuw and Van Driessen (1999). Two commonly chosen fractions are $\gamma = 0.5$ which yields the highest breakdown value (50%) and $\gamma = 0.75$ which gives a better compromise between efficiency and breakdown (25%). To increase efficiency we compute the one-step reweighted MCD (RMCD) defined as

$$t_n = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad \text{and} \quad C_n = c_n \frac{\sum_{i=1}^n w_i (x_i - t_n)(x_i - t_n)^t}{\sum_{i=1}^n w_i}. \quad (1)$$

The weight w_i equals 1 when the squared robust distance $RD^2(x_i) := (x_i - t_n^0)^t (C_n^0)^{-1} (x_i - t_n^0)$ is smaller than the cutoff value $\chi_{p,0.975}^2$. Otherwise the weight equals zero. The factor c_n makes the RMCD consistent and yields more reliable outlier identification (Pison et al. 2002).

S-estimators (t_n, C_n) of location and scatter minimize $\det(C)$ subject to the condition

$$\frac{1}{n} \sum_{i=1}^n \rho(\sqrt{(x_i - t)^t C^{-1} (x_i - t)}) = b \quad (2)$$

with $t \in \mathbb{R}^p$ and C in the class of positive definite symmetric matrices of size p . The constant b equals $E_{F_0} \rho(\|x\|)$, with $F_0 = N_p(0, I)$. For well-chosen ρ functions, S-estimators have a positive breakdown value and bounded influence function (Lopuhaä 1989). The most common choice for the function ρ is the Tukey biweight given by $\rho(y) = \min(\frac{y^2}{2} - \frac{y^4}{2c^2} + \frac{y^6}{6c^4}, \frac{c^2}{6})$

with c determined by $\rho(c) = b/r$ where $0 < r \leq 0.5$ is the breakdown value of the estimator. Ruppert (1992) constructed an efficient algorithm to compute S-estimates.

3 PRINCIPAL COMPONENT ANALYSIS

Given a set of variables X_1, \dots, X_p , in principal component analysis we construct a set of new variables Y_i ($i = 1, \dots, p$) which are uncorrelated linear combinations of the original variables with maximum variance. Principal component analysis is often used to represent the original p -dimensional data in a low dimensional subspace of dimension k spanned by the principal components that have the k largest variances. Denote $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ the eigenvalues of the (theoretical) correlation matrix R and e_1, e_2, \dots, e_p the corresponding eigenvectors. Write $X = [X_1, X_2, \dots, X_p]$, then the principal components are given by $Y_i = e_i^t X$ with $Var(Y_i) = \lambda_i$ and $Corr(Y_i, Y_j) = 0$ $i \neq j$. Hence, the principal components are determined by the eigenvalue-eigenvector pairs of the correlation matrix R .

The influence functions of the eigenvalues $l_i(F)$ and eigenvectors $v_i(F)$ corresponding to the classical correlation matrix $R(F)$ can be derived (from Critchley 1985) and are given by expressions (A1) and (A2) in the appendix. For a given dataset $X_n = \{x_1, \dots, x_n\}$ we obtain the empirical influence functions $EIF(x_i, l_j)$ and $EIF(x_i, v_j)$ for $j = 1, \dots, p$ by replacing the model parameters with the robust estimates obtained from the data. The influence of an observation on an eigenvector is measured by computing the norm of the p -dimensional vector $EIF(x_i, v_j)$ as in Shi (1997).

To summarize the influence of an observation on the analysis, we compute its overall influence on the first k eigenvalues and eigenvectors of interest. The overall influence of an observation is the norm of its empirical influences on each of the k components. We scale this norm to avoid increasing values with increasing dimensions. Formally, the overall influence of observation x_i is given by

$$EIF_k(x_i, l) = \sqrt{\frac{1}{k} \sum_{j=1}^k EIF(x_i, l_j)^2} \quad (3)$$

$$EIF_k(x_i, v) = \sqrt{\frac{1}{kp} \sum_{j=1}^k \sum_{l=1}^p EIF(x_i, v_{jl})^2} \quad (4)$$

where $v_j = (v_{j1}, \dots, v_{jp})^t$ for $j = 1, \dots, p$.

To detect influential points a cutoff value for the overall influences is determined by Monte Carlo simulation. We generate $m=100$ datasets of the same sample size as the original dataset. The datasets are generated from a multivariate normal distribution with the robust correlation matrix estimate of the original dataset as correlation matrix. In this way, these datasets have the same correlation structure as the bulk of the original data. For each of the datasets the overall influence is computed for all data points by substituting the classical parameter estimates in expressions (A1) and (A2). This is reasonable because the datasets follow the model such that the classical estimates are the optimal parameter estimates. Moreover, they can also be computed faster. The cutoff value now becomes the 95 % quantile of the overall influences obtained for the data points of these 100 datasets. In

this way we derive a critical value for the overall influence under the null hypothesis that there are no influential points in the dataset.

The results can be displayed graphically as follows. The overall empirical influence of the observations is plotted versus their robust distance $RD(x_i)$. To this plot we add a horizontal line that corresponds with the cutoff value for the overall influences. Similarly, the vertical line in the plot is a cutoff value for the robust distances, given by the 95% quantile of the robust distances obtained from the Monte Carlo simulation. We thus obtain a critical value for the robust distances under the null hypothesis that there are no outliers in the dataset. We prefer this simulated value over the usual Chi-square quantile because it has been shown that the Chi-square approximation can be very slow (Hardin and Rocke 1999). As a result, the plot is divided into four different regions. The lower left quadrant contains the regular points which are close to the center of the data and have a small influence on the analysis. Non-outlying points that highly influence the estimates are visible in the upper left quadrant. The outliers are identified in the right quadrants of the plots. Highly influential outliers are shown in the upper right quadrant while outliers with only small influence are found in the lower right quadrant. Throughout the paper we will illustrate these plots using 25% and 50% breakdown MCD and S-estimators. In practice, data with more than 20% of outliers do not often occur, so in most situations 25% breakdown estimators suffice. However, sometimes one has to deal with low quality data that require the use of 50% breakdown estimators.

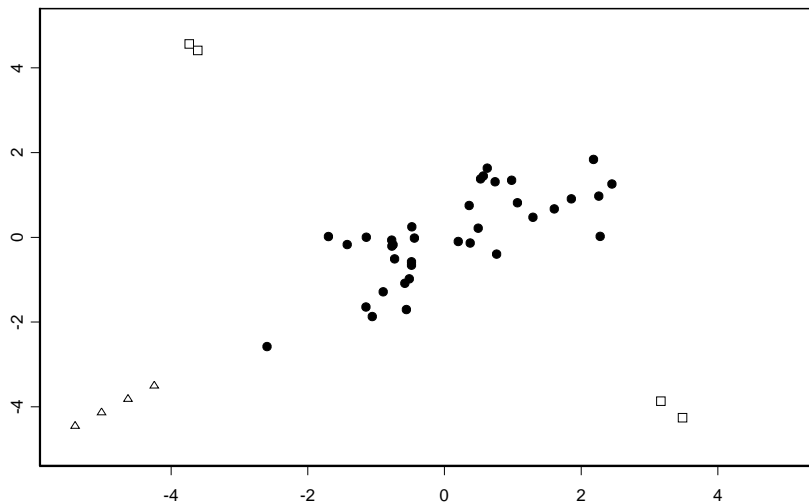


Figure 1: Scatter plot of the generated data with $n = 40$ and $p=2$. Outliers according to first situation are indicated by \triangle and according to second situation by \square .

Let us look at a generated dataset to further clarify these diagnostic plots. We generated 40 points from a bivariate Gaussian distribution with correlation 0.7 between the two variables. First, we moved four objects in the direction of the first principal component (see Figure 1). The diagnostic plots based on RMCD with $\gamma = 0.75$ are shown in Figure 2. From Figure 2a we immediately see that the four outliers lie in the upper right quadrant

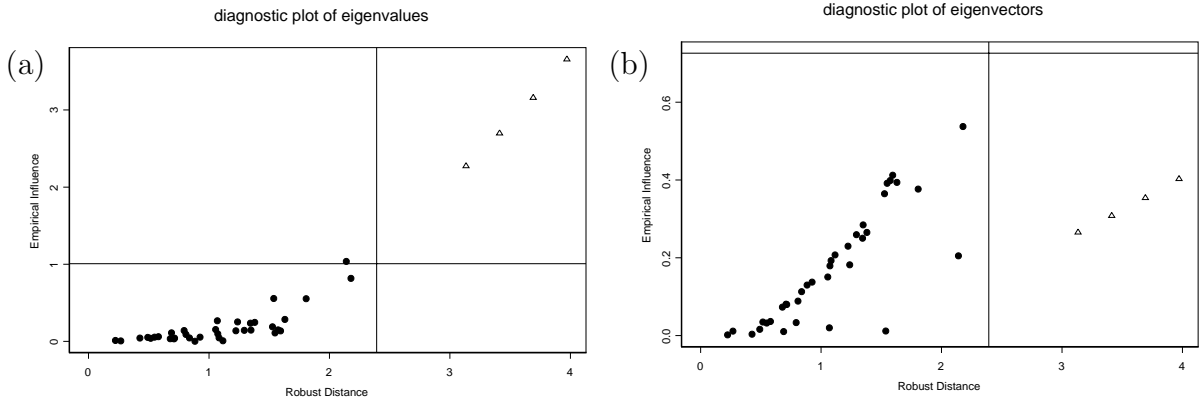


Figure 2: Diagnostic plot based on RMCD for generated data with $p = 2$ and $n = 40$; (a) of the eigenvalues; (b) of the eigenvectors.

meaning that they highly influence the eigenvalue estimates. On the other hand, Figure 2b shows that the outliers hardly influence the eigenvectors. This corresponds with intuition. Moving points in the direction of the first principal component does not affect the directions of the principal components. However, the variance of the first principal component becomes larger due to the outliers as reflected by the high influences in Figure 2a. Note that outliers always affect the classical correlation matrix which results in a high influence on its eigenvalues. For principal component analysis the plot for the eigenvectors (Figure 2b) is thus more informative because it shows whether the outliers also affect the directions of the principal components.

Secondly, we moved the four observations in the direction of the second principal component (see Figure 1). Figure 3 shows the corresponding diagnostic plot for the eigenvectors based on S-estimates ($r = 0.25$). We see that the outliers now also highly influence the estimated eigenvectors. Also here this is intuitively clear. The four outliers increase the variance in the direction of the second principal component so that it becomes larger than the variance in the direction of the first original principal component (see Figure 1). Hence, the outliers do not only influence the eigenvalues but also drastically affect the eigenvectors.

Example. We consider a real data example of Lee (1992). This dataset consists of measurements on the properties of handsheets made from pulp samples. The dataset contains 14 variables on 62 samples. Using RMCD ($\gamma = 0.75$), a robust principal component analysis is performed. Figure 4 shows that there are seven influential outliers for the eigenvectors of which observation 50 is a huge outlier with a very large influence. The influence of the other outliers is much less severe in comparison. A classical principal component analysis of the full dataset would thus be mainly determined by observation 50. Given the number of outliers and their high influence as revealed in Figure 4 further analysis of the data is needed.

Tanaka (1988) studied the influence function of the projection matrix $P_1 P_1^t$ where $P_1 = (e_1, \dots, e_k)$ to investigate the influence of each observation on the subspace spanned by the first k eigenvectors. Using the empirical counterpart of this influence function (expression (A3) in the appendix) we obtain a diagnostic plot that investigates the sensitivity of

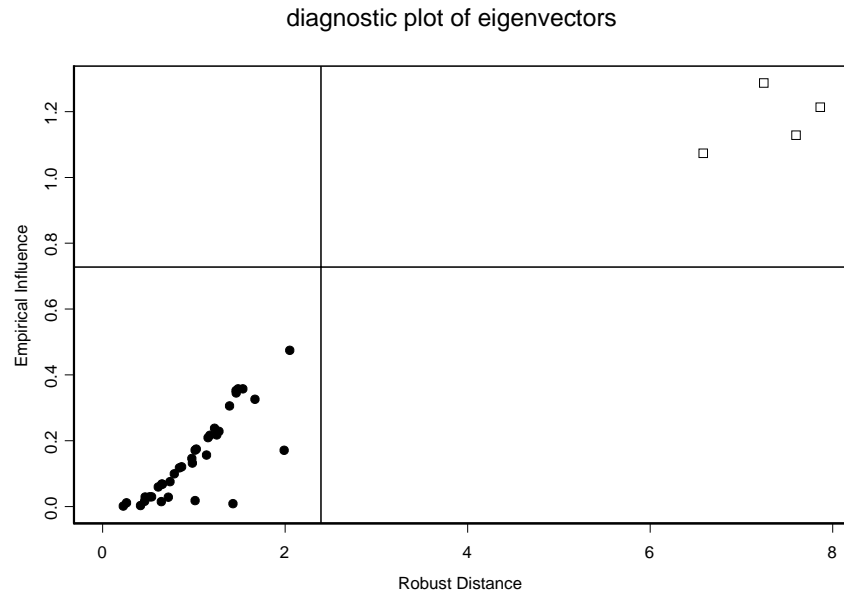


Figure 3: Diagnostic plot of the eigenvectors based on S-estimates for generated data with $p = 2$ and $n = 40$.

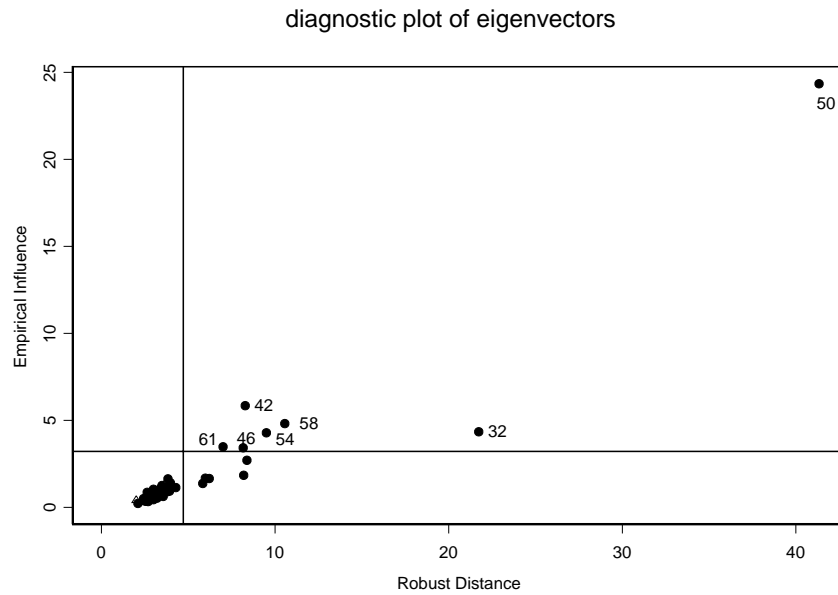


Figure 4: The diagnostic plot of the eigenvectors based on RMCD for the pulp and paper dataset.

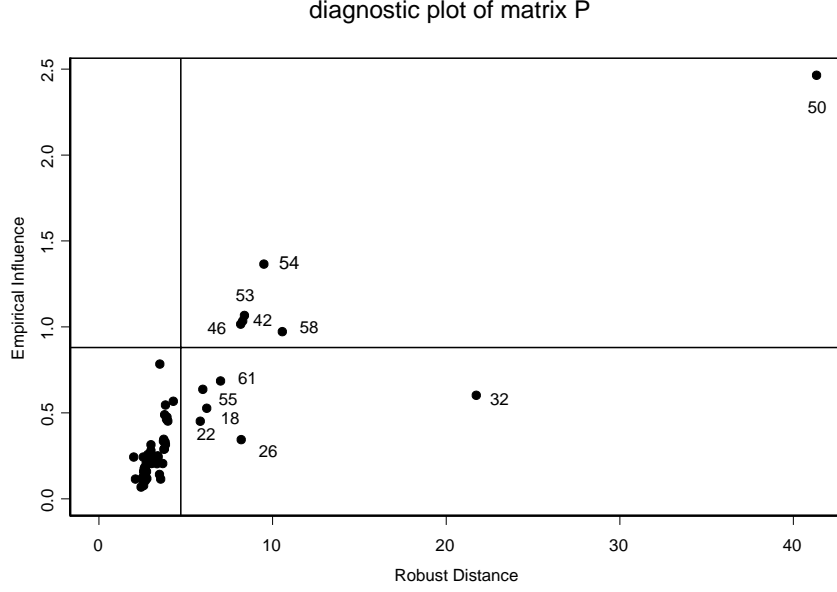


Figure 5: The diagnostic plot based on RMCD of the projection matrix $P_1 P_1^t$ on the subspace spanned by the first three eigenvectors of the pulp and paper data.

the subspace. For example, the robust principal components analysis of the pulp and paper data reports that the first three principal components account for 90% of the total variation. Figure 5 shows the diagnostic plot for the corresponding projection matrix. Again observation 50 has the largest impact on the projection matrix but also the other outliers all have a considerable influence on the projection matrix, so clearly the subspace spanned by the first three principal components cannot be considered to be stable.

4 CANONICAL CORRELATION ANALYSIS

In canonical correlation analysis we consider two groups of variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$. Denote $s = \min(p, q)$. In the first step linear combinations of X as well as Y are taken such that the correlation between the two sets of linear combinations is maximal. In the following steps the pair of linear combinations maximizes the correlation among all pairs uncorrelated with the already determined canonical variables. The purpose of this analysis is to measure the strength of association between the two sets of variables. Let R be the correlation matrix of $Z = (X^t, Y^t)^t$ and partition R as

$$R = \begin{pmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{pmatrix}. \quad (5)$$

Then the s pairs of canonical variables are given by $U_i = e_i^t R_{XX}^{-1/2} X$ and $V_i = f_i^t R_{YY}^{-1/2} Y$ with $\text{Corr}(U_i, V_i) = \rho_i$. Here (e_1, \dots, e_s) are eigenvectors of $R_{XX}^{-1/2} R_{XY} R_{YY}^{-1} R_{YX} R_{XX}^{-1/2}$ and (f_1, \dots, f_s) are eigenvectors of $R_{YY}^{-1/2} R_{YX} R_{XX}^{-1} R_{XY} R_{YY}^{-1/2}$ corresponding to the s largest eigenvalues. The values (ρ_1, \dots, ρ_s) are the square roots of the eigenvalues corresponding to

(e_1, \dots, e_s) . Note that they are also the square roots of the eigenvalues corresponding to (f_1, \dots, f_s) . Hence, the canonical correlation analysis is completely defined by the two sets of eigenvectors together with the corresponding eigenvalues.

The influence functions of the classical estimators $r_i(F)$, $a_i(F)$, and $b_i(F)$ for the parameters ρ_i , $\alpha_i^t = e_i^t R_{XX}^{-1/2}$ and $\beta_i^t = f_i^t R_{YY}^{-1/2}$ can be derived from (Romanazzi 1992) and are given by expressions (A4)-(A6) in the appendix.

As before, to obtain the empirical influences we substitute robust parameter estimates in these influence functions. Overall influences are then computed for the eigenvalues and for the two sets of eigenvectors by taking scaled norms of the components. Finally, we obtain the following three overall empirical influences for each observation

$$EIF(z_i, r) = \sqrt{\frac{1}{s} \sum_{j=1}^s EIF(z_i, r_j)^2} \quad (6)$$

$$EIF(z_i, a) = \sqrt{\frac{1}{s p} \sum_{j=1}^s \sum_{l=1}^p EIF(z_i, a_{jl})^2} \quad (7)$$

$$EIF(z_i, b) = \sqrt{\frac{1}{s q} \sum_{j=1}^s \sum_{l=1}^q EIF(z_i, b_{jl})^2} \quad (8)$$

The cutoff values for these overall influences are computed by Monte Carlo simulation.

For canonical correlation analysis the influence measures (6)-(8) yield three diagnostic plots that display the results of the influence analysis. We now consider a few simple examples to illustrate the plots. In the first example we took $X = X_1$ and $Y = (Y_1, Y_2)^t$. We generated 40 observations of the variable $Z = (X_1, Y_1, Y_2)^t$ according to a multivariate Gaussian distribution with correlation matrix

$$R = \begin{pmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

First, we added five outliers whose correlation between X_1 and Y_1 equals 1 instead of 0.7. A canonical correlation analysis on the original observations would ideally give $U = X_1$ and $V = 1 * Y_1 + 0 * Y_2$ which yields the maximal correlation of 0.7. In the analysis of this dataset the canonical variable U cannot change because X contains only one variable, so $U = X$ and $\alpha = 1$. In this situation, also variable V and thus the canonical vector $\beta = (1, 0)$ are unchanged because the correlation between X_1 and Y_1 increases in the presence of the outliers. Therefore, the diagnostic plots of α and β will give (approximately) zero influence to all observations. The diagnostic plot of the correlations based on RMCD ($\gamma = 0.75$) is given in Figure 6. We see that the five added data points are outliers that highly influence the correlation estimates. Hence, although these five outliers do not influence the canonical variables, they do influence the estimates of the correlation between these canonical variables.

Secondly, we added five outliers with correlation between X_1 and Y_2 equal to 1 instead of 0 for the original observations. As before, α equals 1, but the canonical variable V now

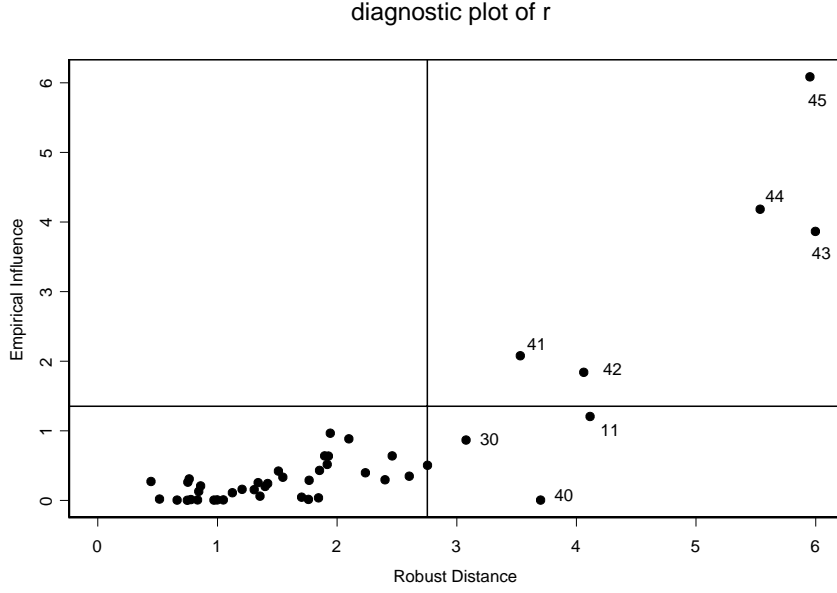


Figure 6: The diagnostic plot based on RMCD of the correlation for generated data with $p = 3$ and $n = 45$.

becomes totally different in the presence of the outliers. Without the outliers we know that ideally $\beta = (1, 0)$. However, since the outliers have a totally different correlation, they affect the analysis leading to a different result ($\beta = (0.98, 0.18)$). This is reflected in the diagnostic plot of β (Figure 7) which clearly shows that the five outliers now also highly influence the estimates of β .

Example. In this real data example a canonical correlation analysis is used to investigate the relationship between three physical measures (weight, waist perimeter, and pulse rate) and three measures of performance on physical exercises (number of pull ups, number of flexures, and number of jumps). The dataset (Tenenhaus 1998) contains measurements of the 6 variables on 20 persons. The diagnostic plots for the canonical correlation analysis based on RMCD ($\gamma = 0.5$) in Figure 8 and Figure 9 show that there are three influential outliers (3, 9, and 10). observation 10 is a huge outlier with a large influence on all estimates. On the other hand observation 9 only affects the canonical variables U_i (Figure 9a) while observation 3 only affects the V_i (Figure 9b).

Romanazzi (1992) also considers the average squared canonical correlation $\psi_t = \frac{1}{t} \sum_j \rho_j^2$ with t the rank of R_{XY} as an index of multivariate association between X and Y . The influence function $IF(x, \psi_t, F)$ is given in (A7) in the appendix. If we consider again the real data example we see that the result now is more conservative. Based on the average squared correlation (Figure 10) only observation 10 is identified as influential outliers while observations 3, 9 and 10 were labeled influential in Figure 8. This was also the case in other examples.

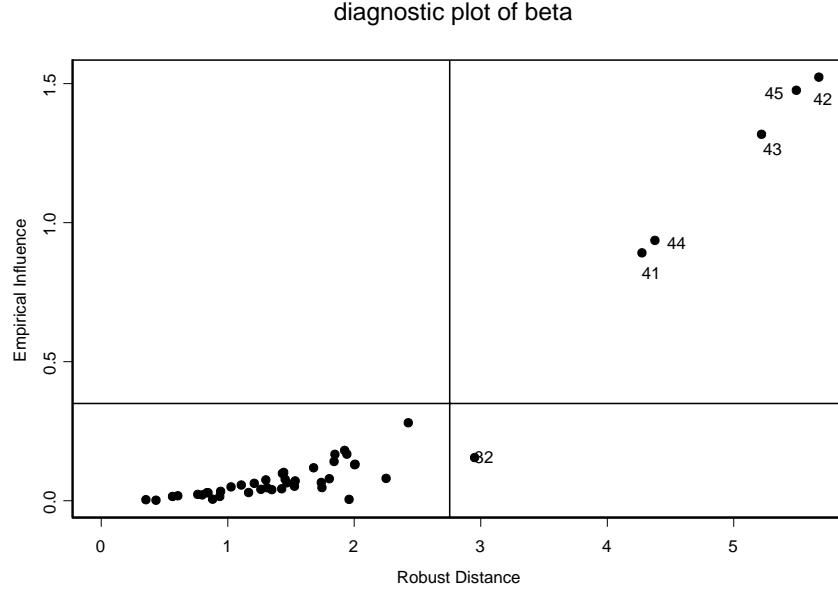


Figure 7: The diagnostic plot of beta based on S-estimates for generated data with $p = 3$ and $n = 45$.

5 PRINCIPAL FACTOR ANALYSIS

In factor analysis we want to approximate the p original variables $X = (X_1, \dots, X_p)^t$ by a smaller number $k \leq p$ of unknown variables $\Phi = (\Phi_1, \dots, \Phi_k)^t$ called factors. These latent factors describe the correlation matrix R of the original variables. In particular, the model assumes that $X - \mu = \Lambda\Phi + \varepsilon$ with $\Lambda \in \mathbb{R}^{p \times k}$ the matrix of factor loadings and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)^t$. The random vectors Φ and ε are assumed to be independent, $E(\Phi) = 0$, $Cov(\Phi) = I$, $E(\varepsilon) = 0$ and $Cov(\varepsilon) = diag(\Psi)$ with $\Psi = (\Psi_1, \dots, \Psi_p)^t$. Under these assumptions we obtain

$$R = \Lambda\Lambda^t + diag(\Psi) \quad (9)$$

The loading matrix Λ is only determined up to an orthogonal transformation.

Note that factor analysis is completely determined by $\Lambda\Lambda^t$ and Ψ . Several estimation methods exist such as the principal components solution, principal factor analysis, maximum likelihood estimation and iteratively reweighted least squares. Pison et al (2003) have shown that of the two frequently used methods principal factor analysis and maximum likelihood the former works best with robust estimates. Therefore, we consider principal factor analysis and denote the estimators of Λ and Ψ obtained from the classical correlation matrix by $L(F)$ and $P(F)$. The influence functions of $L(F)L(F)^t$ and $P(F)$ have been derived in Tanaka and Odaka (1989) and Pison et al. (2003). See expressions (A8)-(A13) in the appendix.

Substituting robust parameter estimates in expressions (A8)-(A13) and solving the resulting system of equations yields the empirical influences of LL^t and P . Finally, we consider

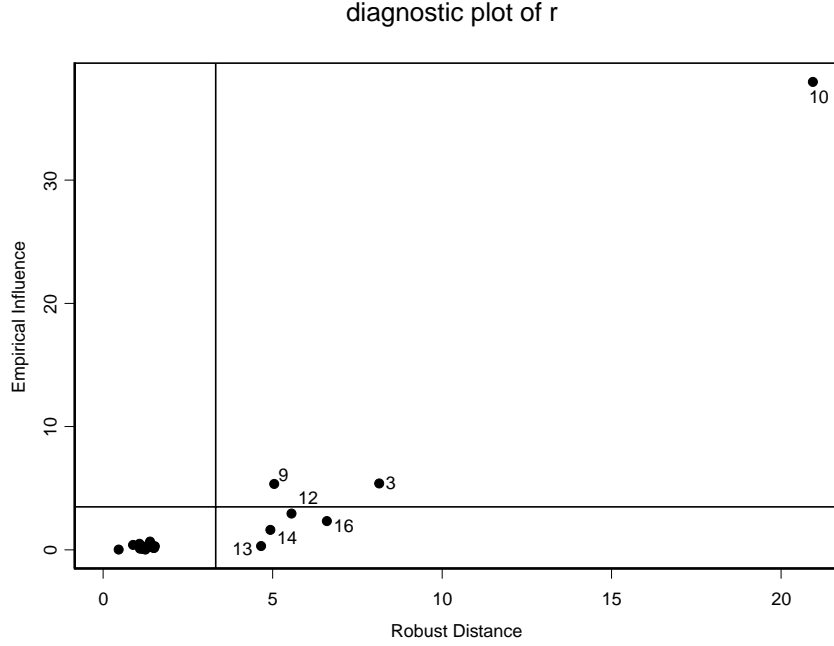


Figure 8: The diagnostic plot of the canonical correlation based on RMCD estimates for the Physical dataset with $p = 6$ and $n = 20$.

the overall empirical influence

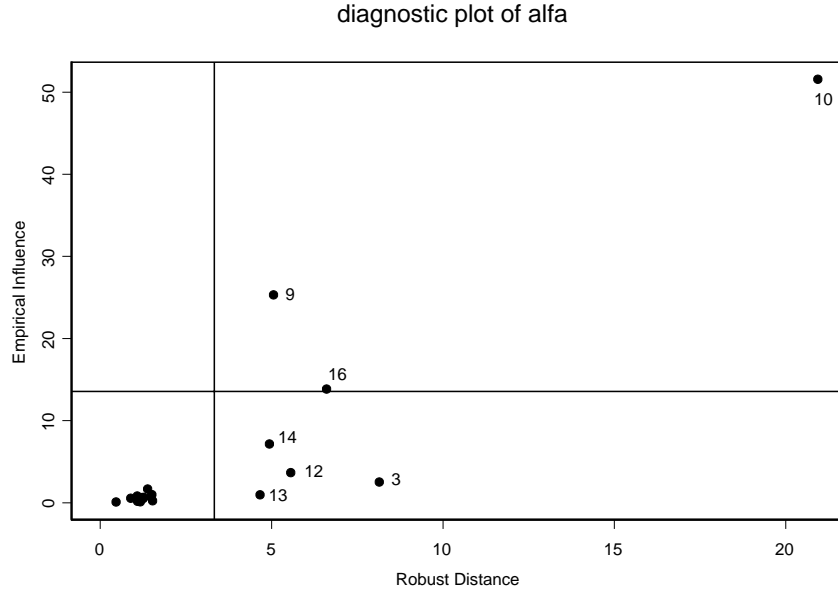
$$EIF_k(x_i, LL^t) = \sqrt{\frac{1}{p^2} \sum_{j=1}^p \sum_{l=1}^p EIF(x_i, LL_{jl}^t)^2} \quad (10)$$

Since the diagonal elements of the correlation matrix are fixed, it follows from (9) that the influence on the components of P is the same (except for the sign) as the influence on the diagonal elements of LL^t . Therefore, expression (10) above also contains the effect on P . If interest is only in reproducing the correlations and the specific variances Ψ are considered nuisance, then the mean over all off-diagonal elements of LL^t should be taken in (10).

As an example, we generated 50 observations according to a factor model with loading matrix and specific variances given in Table 1. Note that the corresponding correlation matrix is given by (9). We added 15 observations generated from a distribution with the same correlation matrix but with higher variances implying that some of these points will be outlying. The diagnostic plot of a principal factor analysis based on RMCD ($\gamma = 0.5$) in Figure 11 clearly shows that none of the points highly influence the principal factor estimates as expected.

Example. Here, we analyze the Swiss bank notes data (Flury and Riedwyl 1988). This dataset contains measurements of 5 properties on 100 forged bank notes of 1000 Swiss francs. The diagnostic plot for principal factor analysis with $k = 2$ factors using RMCD ($\gamma = 0.75$) is shown in Figure 12. From the 16 outliers detected by RMCD, 11 points are identified as influential for the factor analysis. This is further illustrated in Table 2. The first column

(a)



(b)

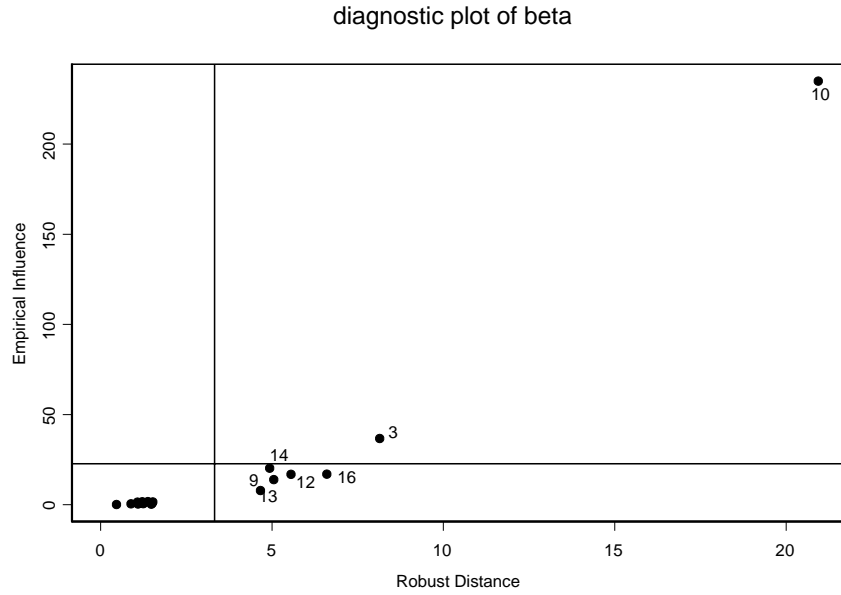


Figure 9: The diagnostic plot based on RMCD estimates for the Physical dataset with $p = 6$ and $n = 20$; (a) of the canonical vectors α ; (b) of the canonical vectors β .

in this table gives the loadings obtained from classical factor analysis applied to the dataset without all the outliers. The estimates in the second column are obtained by only excluding the influential outliers in Figure 12. The third column contains the loadings based on the whole dataset. Note that the loadings in the third column are totally different from those in the first and second column which confirms that the analysis of the complete data has been

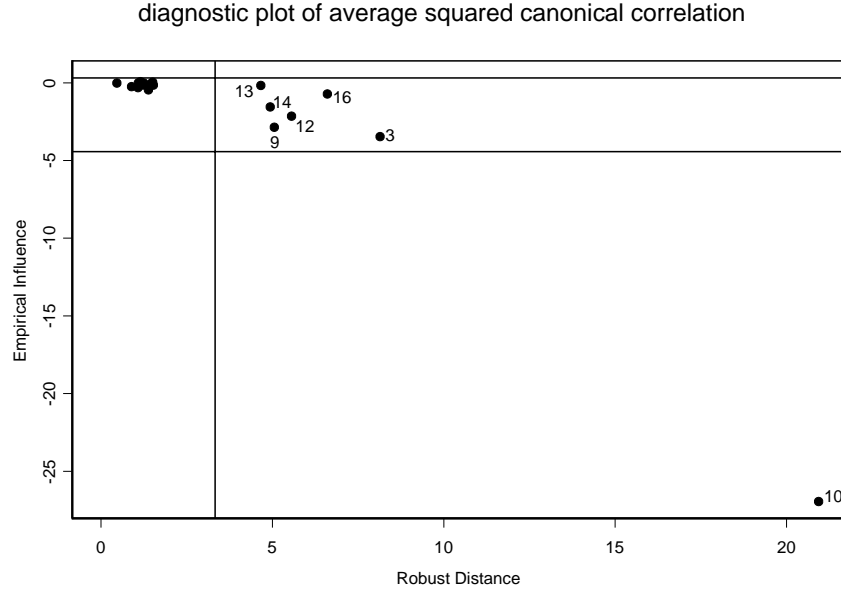


Figure 10: The diagnostic plot based on RMCD estimates for the Physical dataset with $p = 6$ and $n = 20$ of the average squared canonical correlation.

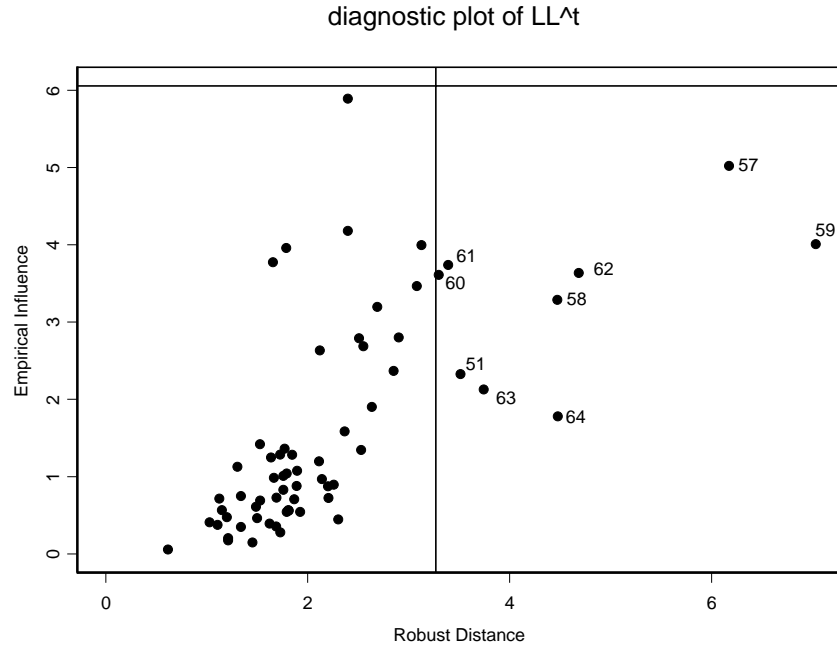


Figure 11: The diagnostic plot of LL^t based on RMCD for generated data with $p = 5$ and $n = 65$.

Table 1: The loading matrix and the specific variances of the 50 observations.

	Λ_1	Λ_2	Ψ
X1	0.783	-0.217	0.34
X2	0.773	-0.458	0.193
X3	0.794	-0.234	0.315
X4	0.713	0.472	0.269
X5	0.712	0.524	0.219

affected by the influential outliers. As expected, the first two analyses give similar results for the loadings which shows that the diagnostic plot indeed separates influential outliers from non-influential outliers.

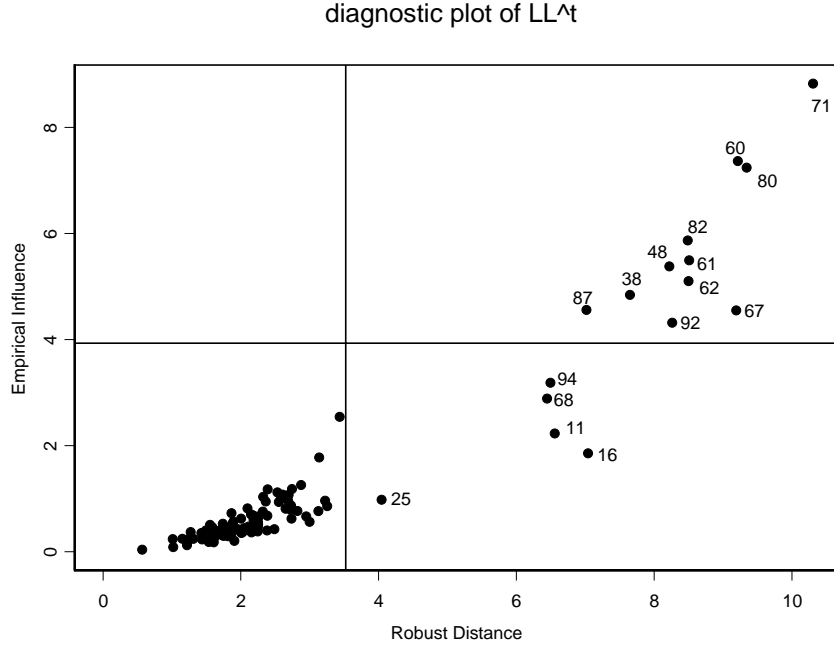


Figure 12: The diagnostic plot based on RMCD for the Swiss dataset with $p = 5$ and $n = 100$; (a) of LL^t .

6 DISCUSSION

We constructed diagnostic plots that summarize in a comprehensive way the influence of the observations on a multivariate data analysis. These plots are based on robust distances of the observations and their empirical influences on the parameter estimates for the multivariate

Table 2: The classical loading estimates based on the regular points, the data without influential outliers and the entire dataset.

length bill	0.554	0.212	0.462	0.307	-0.143	0.403
height left	0.785	-0.180	0.800	-0.104	0.000	0.807
height right	0.769	0.000	0.795	0.000	0.109	0.744
distance frame-bottom	0.000	-0.969	0.000	-0.899	0.974	-0.199
distance frame-top	0.000	0.846	0.000	0.814	-0.664	0.000
length diagonal	0.620	0.223	0.454	0.000	0.302	0.000

model. As a result we can identify regular points, non-outlying influential points, outliers that are not influential and influential outliers and as such gain more insight in the data.

These plots are meant to complement other graphical displays depicting the results of the analysis. For example, in principal components analysis it is instructive to make a scatterplot matrix of the scores. In fact, using different colors and plotting symbols, outliers and influential points can also be highlighted in these plots. Figure 13 shows the scatterplot of the scores for the first three principal components of the pulp-paper data obtained from the robust analysis. Different symbols identify nonoutlying influential points (■) and influential outliers (▲). In this example all outliers were influential. This plot shows which principal components will be most affected by the influential points. Similar plots can be constructed for the scores in canonical correlation and factor analysis. Outliers and influential points can also be highlighted in other useful plots in multivariate data analysis such as biplots (Gower and Hand 1996).

We based the detection of influential points and outliers on marginal tests that assume under the null hypothesis respectively that the data contain no influential points and no outliers. The significance level for the joint null hypothesis that an observation is non-outlying and non-influential will be higher than the 5% significance level we used in both marginal tests. This significance level can easily be obtained by determining the percentage of irregular points in the Monte-Carlo simulation. From Table 3 we see that the significance level for the joint test is around 9% in all cases which still is acceptable for exploratory data analysis. The results suggest that a Bonferroni correction might be appropriate to obtain a 5% significance level for the joint testing procedure. This is confirmed by the last column in Table 3 whose values are all close to the nominal value .05. Since the significance level for the joint test is obtained as a result in the Monte-Carlo simulation, other (problem-driven) combinations of the significance levels for the marginal tests can be combined to get the desired joint significance level.

The plots in this paper have been constructed using robust distances obtained from high breakdown, bounded influence location and scatter estimators to detect outliers and empirical influences based on robust estimates to measure the influence of the observations. In general, however, this type of plots can be constructed for any choice of outlier detection rule and influence measure available.

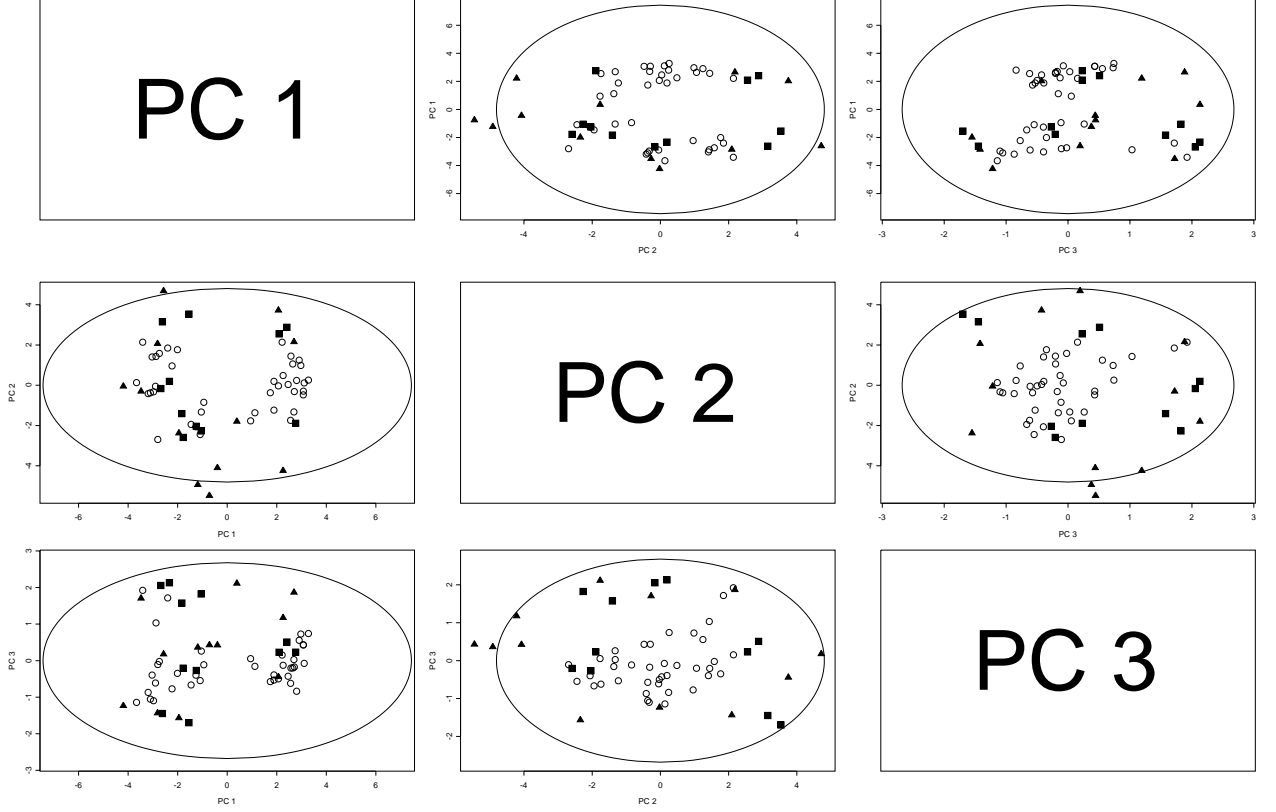


Figure 13: Scatterplot matrix of the scores for the first three principal components of the pulp and paper data. Nonoutlying influential points (■) and influential outliers (▲) are identified.

7 APPENDIX

Principal component analysis. The influence functions of the eigenvalues $l_i(F)$ and eigenvectors $v_i(F)$ corresponding to the classical correlation matrix $R(F)$ are given by Critchley (1985)

$$IF(x, l_j, F) = \tilde{z}_j^2 - \lambda_j e_j^t D_{\tilde{x}} e_j \quad (\text{A1})$$

$$IF(x, v_j, F) = \sum_{\substack{k=1 \\ k \neq j}}^p (\tilde{z}_k \tilde{z}_j - \frac{\lambda_k + \lambda_j}{2} e_j^t D_{\tilde{x}} e_k) \frac{e_k}{\lambda_j - \lambda_k}. \quad (\text{A2})$$

Here $\tilde{z}_j = e_j^t \tilde{x}$ and $D_{\tilde{x}} = \text{diag}(\tilde{x} \tilde{x}^t)$ where $\tilde{x} = \Sigma_D^{-1/2}(x - \mu)$ with $\Sigma_D = \text{diag}(\Sigma)$ the diagonal matrix with the same diagonal elements as Σ .

From Tanaka (1988) we obtain that the influence functions of $P_1 P_1^t$ is given by

$$IF(x, P_1, P_1^t, F) = \sum_{s=1}^k \sum_{r=k+1}^p (\lambda_s - \lambda_r)^{-1} (e_s^t IF(x, R, F) e_r) (e_s e_r^t + e_r e_s^t) \quad (\text{A3})$$

Table 3: Significance level for the joint null hypothesis that an observation is non-outlying and non-influential obtained by Monte-Carlo simulation. Marginal levels are $\alpha = 5\%$ and $\alpha = 2.5\%$.

Model	Dataset	parameter	$\alpha = 5\%$	$\alpha = 2.5\%$
PCA	pulp-paper	v	0.093	0.048
		$P_1 P_1^t$	0.088	0.045
CCA	physical performance	ρ	0.089	0.048
		α	0.092	0.049
		β	0.094	0.049
		ψ_q	0.091	0.046
PFA	Swiss banknotes	LL^t	0.094	0.049

Canonical correlation analysis. The influence functions of the classical estimators $r_i(F)$, $a_i(F)$, and $b_i(F)$ for the parameters ρ_i , $\alpha_i^t = e_i^t R_{XX}^{-1/2}$ and $\beta_i^t = f_i^t R_{YY}^{-1/2}$ can be derived from (Romanazzi 1992) and are given by

$$IF(z, r_j, F) = u_j v_j - \frac{1}{2} \rho_j u_j^2 - \frac{1}{2} \rho_j v_j^2 \quad (\text{A4})$$

$$IF(z, a_j, F) = \left[\sum_{\substack{k=1 \\ k \neq j}}^s \left(\frac{\rho_j (v_j - \rho_j u_j) u_k + \rho_k (u_j - \rho_j v_j) v_k}{\rho_j^2 - \rho_k^2} \right) \alpha_k \right] + \frac{1}{2} (D_{\tilde{x}} \alpha_j - u_j^2 \alpha_j) \quad (\text{A5})$$

$$IF(z, b_j, F) = \left[\sum_{\substack{k=1 \\ k \neq j}}^s \left(\frac{\rho_j (u_j - \rho_j v_j) v_k + \rho_k (v_j - \rho_j u_j) u_k}{\rho_j^2 - \rho_k^2} \right) \beta_k \right] + \frac{1}{2} (D_{\tilde{y}} \beta_j - v_j^2 \beta_j) \quad (\text{A6})$$

where $u_k = \bar{\alpha}_k^t (x - \mu_x)$, $\bar{\alpha}_k = D_X^{-1/2} \alpha_j$, $v_k = \bar{\beta}_k^t (y - \mu_y)$, $\bar{\beta}_k = D_Y^{-1/2} \beta_j$, $D_{\tilde{x}} = \text{diag}(\tilde{x} \tilde{x}^t)$, $D_{\tilde{y}} = \text{diag}(\tilde{y} \tilde{y}^t)$, $D_X = \text{diag}(\Sigma_{XX})$, $D_Y = \text{diag}(\Sigma_{YY})$, $\tilde{x} = \text{diag}(\Sigma_X)^{-1/2} (x - \mu_x)$ and $\tilde{y} = \text{diag}(\Sigma_Y)^{-1/2} (y - \mu_y)$. From Romanazzi (1992) we have that the influence functions of $\psi_t = t^{-1} \sum_j \rho_j^2$ is given by

$$IF(\psi_t) = \frac{1}{t} \sum_j 2 \rho_j IF(x, r_j, F) \quad (\text{A7})$$

Principal factor analysis. The influence functions of the functionals $L(F)L(F)^t$ and $P(F)$ at $F = N_p(\mu, \Sigma)$ can be obtained from Tanaka and Odaka (1989), Pison et al. (2003). Let us denote $(\lambda_1, \dots, \lambda_k)$ and (e_1, \dots, e_k) for the eigenvalues and eigenvectors of $\Lambda \Lambda^t$. We write $l_1(F), \dots, l_k(F)$ and $v_1(F), \dots, v_k(F)$ for the functionals of the corresponding classical estimators which are thus the eigenvalues and eigenvectors of $L(F)L(F)^t$. Then the influence

functions are given by the following expressions.

$$IF(x, l_j, F) = e_j^t [(IF(x, R, F) - \text{diag}(IF(x, P, F)))] e_j \quad (\text{A8})$$

$$\begin{aligned} IF(x, v_j, F) &= \sum_{\substack{l=1 \\ l \neq j}}^k \frac{1}{\lambda_l - \lambda_j} \{e_l^t [-IF(x, R, F) + \text{diag}(IF(x, P, F))]\} e_j e_l \\ &+ \sum_{l=k+1}^p \frac{-1}{\lambda_j} \{a_l^t [\text{diag}(IF(x, P, F)) - IF(x, R, F)] e_j\} a_l. \end{aligned} \quad (\text{A9})$$

where (a_{k+1}, \dots, a_p) form an orthonormal basis of the orthogonal complement (e_1, \dots, e_p) . Furthermore,

$$IF(x, P_j, F) = IF(x, R_{jj}, F) - \sum_{l=1}^k IF(x, l_l, F) e_{l_j}^2 - \sum_{l=1}^k 2\lambda_l e_{l_j} IF(x, v_{l_j}, F) \quad (\text{A10})$$

$$\begin{aligned} IF(x, LL^t, F) &= \sum_{j=1}^k \{IF(x, l_j, F) e_j e_j^t + \lambda_j IF(x, v_j, F) e_j^t \\ &+ \lambda_j e_j IF(x, v_j, F)^t\}. \end{aligned} \quad (\text{A11})$$

with

$$\begin{aligned} IF(x, R, F) &= \Sigma_D^{-1/2} IF(x, S, F) \Sigma_D^{-1/2} - \frac{1}{2} \Sigma_D^{-1} IF(x, S_D, F) R \\ &- \frac{1}{2} R \Sigma_D^{-1} IF(x, S_D, F) \end{aligned} \quad (\text{A12})$$

and

$$IF(x, S, F) = (x - \mu)(x - \mu)^t - \Sigma. \quad (\text{A13})$$

Here $S(F)$ is the functional corresponding to the classical estimator of the parameter Σ and as before Σ_D consists of the diagonal of Σ and zeros elsewhere.

ACKNOWLEDGEMENT

We would like to thank the referees for their helpful comments and suggestions.

REFERENCES

- Atkinson, A.C. and Riani, M. (2000), *Robust Diagnostic Regression Analysis*, Springer, Berlin.
- Campbell, N.A. (1978), "The Influence Function as an Aid in Outlier Detection in Discriminant Analysis," *Applied Statistics*, 27, 251-258.
- Cook, R.D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, London.

- Critchley, F. (1985), "Influence in Principal Component Analysis", *Biometrika*, 72, 627–636.
- Croux, C. and Haesbroeck, G. (2000), "Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies," *Biometrika*, 87, 603-618.
- Croux, C. and Dehon, C. (2001), "Analyse Canonique Basée sur des Estimateurs Robustes de la Matrice de Covariance," *La Revue de Statistique Appliquée*, 2, 5-26.
- Davies, L. (1987), "Asymptotic Behavior of S-estimators of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, 15, 1269-1292.
- Fernholz, L.T. (1983), *Von Mises Calculus for Statistical Functionals*, Springer, Berlin.
- Flury, B. and Riedwyl, H. (1988), *Multivariate Statistics: A Practical Approach*, Cambridge University Press.
- Fox, J. (1991), *Regression Diagnostics*, Sage Publications, CA.
- Gnanadesikan, R. (1977), *Methods of Statistical Data Analysis of Multivariate Observations*, Wiley, New York.
- Gower, J.C., and Hand, D.J. (1996), *Biplots*, Chapman and Hall, New York.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust statistics: the approach based on influence functions*, John Wiley and Sons, New York.
- Hardin, J. and Rocke, D.M. (1999), "The Distribution of Robust Distances," Technical Report, University of California at Davis, <http://www.cipic.ucdavis.edu/~dmrocke/preprints.html>
- Jaupi, L., and Saporta, G. (1993), "Using the Influence Function in Robust Principal Components Analysis," in *New Directions in Statistical Data Analysis and Robustness*, eds. S. Morgenthaler, E. Ronchetti, and W.A. Stahel, Basel: Birkhäuser, 147–156.
- Johnson, R.A. and Wichern, D.W. (1998), *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice Hall, New Jersey.
- Kent, J.T. and Tyler, D.E. (1996), "Constrained M-estimation for Multivariate Location and Scatter," *The Annals of Statistics*, 24, 1346-1370.
- Lee (1992), "Relationships between Properties of Pulp-Fibre and Paper," Phd thesis, University of Toronto.
- Lopuhaä, H.P. (1989), "On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance," *The Annals of Statistics*, 17, 1662-1683.
- Lu, J., Ko, D., and Chang, T., "The Standardized Influence Matrix and Its Applications," *Journal of the American Statistical Association*, 92, 1572-1580.
- Maronna, R.A. (1976), "Robust M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51-67.

- Maronna, R.A. and Yohai, U.J. (1998), "Robust Estimation of Multivariate Location and Scatter," in *Encyclopedia of Statistical Sciences Update Volume 2*, (S. Rotz, C. Read, and D. Banks, eds.) Wiley, New York, pp. 589-596.
- Pison G., Rousseeuw P.J., Filzmoser P. and Croux C. (2003), "Robust Factor Analysis," *Journal of Multivariate Analysis*, 84, 145-172.
- Pison, G., Van Aelst, S., and Willems, G. (2002), "Small Sample Corrections for LTS and MCD," *Metrika*, 55, 111-123.
- Romanazzi, M. (1992), "Influence in Canonical Correlation Analysis", *Psychometrika*, 57, 237-259.
- Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P.J. (1985), "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications*, Vol. B, eds. W. Grossmann, G. Pflug, I. Vincze and W. Wertz, Dordrecht: Reidel, 283-297.
- Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, Wiley-Interscience, New York.
- Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212-223.
- Rousseeuw, P.J. and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633-651.
- Ruppert, D. (1992), Computing S-estimators for regression and multivariate location/dispersion, *Journal of Computational and Graphical Statistics* 1, 253-270.
- Shi, L. (1997), "Local Influence in Principal Components Analysis," *Biometrika*, 84, 175-186.
- Tanaka, Y. (1988), "Sensitivity Analysis in Principal Component Analysis: Influence on the subspace spanned by Principal Components," *Communications in Statistics - Theory and Methods*, 17, 3157-3175.
- Tanaka, Y. and Odaka, Y. (1989), "Influential Observations in Principal Factor Analysis," *Psychometrika*, 54, 475-485.
- Tenenhaus, M. (1998), *La Régression PLS. Théorie et Pratique*, Paris:Technip.